Penalized matrix-variate model-based clustering

European Conference on Data Analysis 2022

Alessandro Casa

Joint work with: A.Cappozzo & M.Fop

Faculty of Economics and Management

Free University of Bozen-Bolzano

alessandro.casa@unibz.it

unibz

Napoli, 14th September 2022

What are we dealing with?

- Framework
  $\longrightarrow$ matrix-variate, or three-way, data are increasingly common in different fields
  $\longrightarrow$ multiple variables measured on a set of units in different occasions
  Some examples:
  - Longitudinal data with multiple features
  - Spatio-temporal or multivariate spatial data
- Complex structure, need to account for three layers

Idea

Resort to clustering strategies to uncover parsimonious patterns in these data

Matrix Gaussian mixture models

- Standard Gaussian mixtures are not adequate therefore we resort to Matrix Gaussian mixture models (MGMM)
  $\longrightarrow$ natural three-way data generalization
- Let $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be a set of n matrices, with $\mathbf{X}_i \in \mathbb{R}^{p \times q}$. According to MGMM the density for $\mathbf{X}_i$ is expressed as

$$f(\mathbf{X}_i; \Theta) = \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k)$$

  - $\phi_{p\times q}(\cdot, \mathbf{M}_k, \Omega_k, \Gamma_k)$, $p \times q$ matrix normal distribution
  - τ_k, mixing proportions
  - M_k, k-th component mean matrix
  - Ω_k and Γ_k rows and columns precision matrices
- Set of parameters $\Theta = \{\tau_k, M_k, \Omega_k, \Gamma_k\}_{k=1}^{K}$

Overparameterization on steroids

- Major limitation: $|\Theta|$ can be huge, as it scales quadratically with both p and q
  - Virtually useless even with moderate dimensions
  - Difficult to interpret relationships among variables/occasions across different clusters

- The problem is encountered even with standard GMM where different workarounds have been proposed
  - Constrained modelling
  - Variable selection
  - Sparse estimation

What's out there?

- Possible solutions in the matrix-variate clustering framework (coherent with the two-way taxonomy)
  - Sarkar et al. (2020): eigendecomposition of the component covariance matrices
  - Wang & Melnykov (2020): stepwise variable selection via BIC values comparison

- These approaches induce parsimony in a rigid way, with structures being constant across groups

Work starting point

Assume that all the matrices in Θ possess their own cluster-dependent degrees of sparsity

Sparse matrix-variate mixture model

- We adopt a penalized likelihood approach by maximizing

$$\ell_P(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k) - p_\lambda(\Theta)$$

  - $p_\lambda(\Theta)$, penalty term to be defined
  - $\lambda = (\lambda_1, \lambda_2, \lambda_3)$, vector of penalty coefficients

- Advantages of this approach
  - Reduced number of parameters
  - Cluster-wise conditional independence patterns
  - Easier interpretation of the associations

Choosing the penalty

- Two different specifications for $p_\lambda(\Theta)$
  - lasso + graphical lasso

$$\sum_{k=1}^{K} \lambda_1 ||P_1 \circ M_k||_1 + \sum_{k=1}^{K} \lambda_2 ||P_2 \circ \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||P_3 \circ \Gamma_k||_1$$

  - group lasso + graphical lasso

$$\sum_{k=1}^{K} \lambda_1 \sum_{r=1}^{p} ||m_{r\cdot,k}||_2 + \sum_{k=1}^{K} \lambda_2 ||P_2 \circ \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||P_3 \circ \Gamma_k||_1$$

- P_1, P_2, P_3 matrices with non-negative entries, $m_{r\cdot,k}$ the r-th row of M_k, $||A||_1 = \sum_{jh} |A_{jh}|$ and $||\cdot||_2$ the Euclidean norm

A bit of interpretation

- Penalty on the mean
  - lasso provides element-wise penalization
  - group lasso allows to perform variable selection by setting entire rows of M_k to zero
- Penalty on the precisions
  - Connection with Gaussian graphical models allows for nice visualization and interpretation
  - Chance to resort to *mix & match* approaches thanks to the connection with Gaussian covariance graph models

- Matrices P_1, P_2, P_3 potentially introduce an higher degree of flexibility, with the chance to include prior beliefs

Estimation strategy

- EM-algorithm to maximize the penalized complete log-likelihood

$$\ell_P^c(\Theta; \mathbf{X}) \propto \sum_{i,k} z_{ik} \Big[\log \tau_k + \frac{q}{2} \log |\Omega_k| + \frac{p}{2} \log |\Gamma_k| + \\ -\frac{1}{2} \mathrm{tr}\left\{ \Omega_k (\mathbf{X}_i - M_k) \Gamma_k (\mathbf{X}_i - M_k)^\intercal \right\} \Big] - p_\lambda(\Theta)$$

- E-step $\longrightarrow$ standard updating formula
- M-step $\longrightarrow$ partial optimization strategy
  - τ_k, standard update
  - Ω_k and Γ_k, estimated via suitable modification of the coordinate descent graphical lasso algorithm
  - M_k $\longrightarrow$ cell-wise coordinate ascent (if lasso)
    $\longrightarrow$ proximal gradient descent (if group lasso)

Model selection

- Need to select $\lambda_1, \lambda_2, \lambda_3$ and K.
  It still represents somehow an open problem as exhaustive grid searches are computationally unfeasible

- Some ideas currently on the table
  - conditional search
  - genetic algorithm
  - E-MS algorithm

- Here every suggestion is more than welcome

Some results - Crime data

- `crime data`
  available in the package `MatTransMix` (Zhu et al., 2022), analyzed in Melnykov & Zhu (2019)
- Crime frequency and rate records between 2000 and 2012 ($q = 13$) for $n = 236$ cities in the US.
  Measured variables ($p = 7$)
  `Violent crimes`
  - murder, rape, robbery, aggravated assault

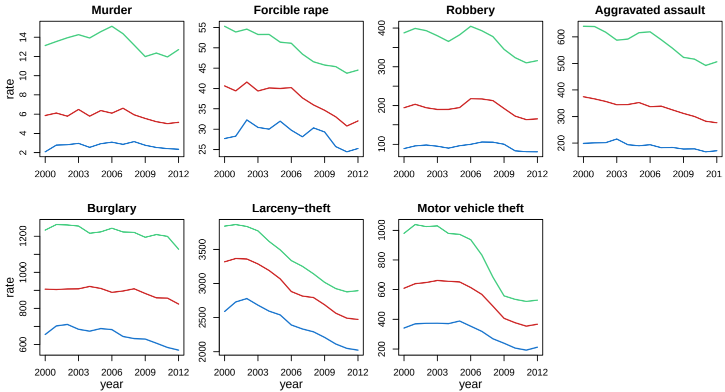
  `Property crimes`
  - motor vehicle theft, burglary, larceny-theft

Aim

Exploit cluster analysis to uncover common time trends for the considered cities

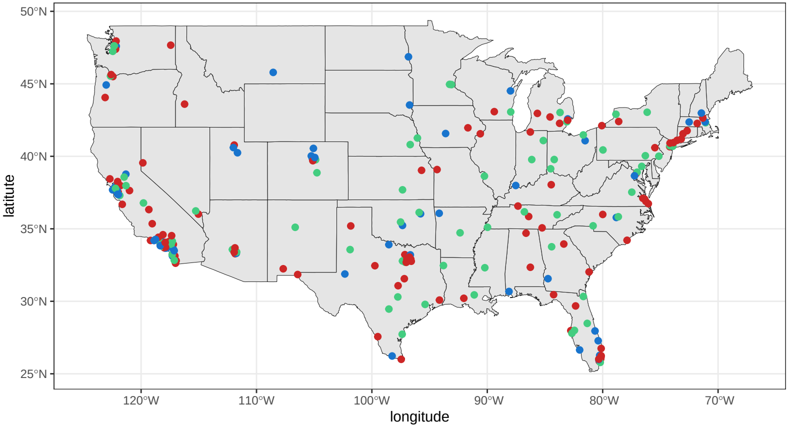
Some results - Crime data

- We obtain $K = 3$ clusters
  - `green`, larger population, highest crime rates
  - `red`, medium population, medium crime rates
  - `blue`, medium population, lowest crime rates
- Banded precision matrices $\longrightarrow$ autoregressive structure



Some results - Crime data

- Indications closed to the ones in Melnykov & Zhu (2019)
  - Eastern USA more dangerous
  - Something along Mississippi belt
  - Large cities are more dangerous then their surroundings



Conclusions & Discussion

- We propose different penalized strategies in the matrix-variate model-based clustering framework

  $\longrightarrow$ different penalties more adequate for different settings
  $\longrightarrow$ easier interpretation of the time/variable relations
  $\longrightarrow$ flexible way to induce parsimony

- Future steps:

  $\longrightarrow$ come up with more clever model selection strategies
  $\longrightarrow$ thorough comparison with potential competitors and alternatives

Some references

Cappozzo, A., Casa, A., & Fop, M. (202x).
Variable selection for matrix-variate model-based clustering via penalized estimation.
Soon on arXiv.

Other relevant references

- Sarkar, S., Zhu, X., Melnykov, V. & Ingrassia, S. (2020). On parsimonious models for modeling matrix data. *Comput Stat Data An*, 142:106822.
- Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J R Stat Soc B*, 68(1): 49-67.
- Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Stat Comput*, 21(4): 511-522.
- Zhou, H., Pan, W. & Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron J Stat*, 3: 1732-1496.